UIT-ViCoV19QA: A Dataset for COVID-19 Community-based Question Answering on Vietnamese Language

Triet Minh Thai
University of Information Technology
VNU-HCM, Vietnam
`19522397@gm.uit.edu.vn`

Ngan Ha-Thao Chu
University of Information Technology
VNU-HCM, Vietnam
`19521882@gm.uit.edu.vn`

Anh Tuan Vo
University of Information Technology
VNU-HCM, Vietnam
`19521226@gm.uit.edu.vn`

Son T. Luu
University of Information Technology
VNU-HCM, Vietnam
`sonlt@uit.edu.vn`

Abstract

For the last two years, from 2020 to 2021, COVID-19 has broken disease prevention measures in many countries, including Vietnam, and negatively impacted various aspects of human life and the social community. Besides, the misleading information in the community and fake news about the pandemic are also serious situations. Therefore, we present the first Vietnamese community-based question answering dataset for developing question answering systems for COVID-19 called UIT-ViCoV19QA. The dataset comprises 4,500 question-answer pairs collected from trusted medical sources, with at least one answer and at most four unique paraphrased answers per question. Along with the dataset, we set up various deep learning models as baseline to assess the quality of our dataset and initiate the benchmark results for further research through commonly used metrics such as BLEU, METEOR, and ROUGE-L. We also illustrate the positive effects of having multiple paraphrased answers experimented on these models, especially on Transformer - a dominant architecture in the field of study.

1 Introduction

Community-based Question answering (CQA) is a task of question answering based on the wisdom of the crowd (Zhang et al., 2021). In CQA, information seekers post their questions on a public website or forum, and other users answer them. This kind of question-answering behavior is popular in people's daily basis. For example, Quora[1] and Reddit[2] are several large forums for question-answering. The CQA enables people to ask and answer questions easily (Qiu and Huang, 2015). Additionally, the development of question and answering systems means computers can now understand and answer the questions of users.

In Vietnam, the year from 2020 to 2021 witnessed the COVID-19 pandemic. Information about the COVID-19 spreading situation, medical care, self-quarantine, vaccination policies, and regulations by the government to prevent the spread of COVID-19 are essential to citizens. People frequently ask questions about the COVID-19 situation, what to do when contacting COVID-19 patients, the vaccination policies, and more. This is our motivation to construct a dataset to help build a question answering system based on CQA about COVID-19 in Vietnamese. Apart from the dataset, we also propose various baseline models to evaluate our dataset's quality. This paper has three main contributions summarized as follows:

1. We introduce UIT-ViCoV19QA, the first community-based question answering collection about the COVID-19 pandemic for Vietnamese constructed from trusted sources. The dataset comprises 4,500 question-answer pairs and is extended to have up to four unique paraphrased answers per question through an efficient paraphrasing process.

2. We assess the dataset's quality and estab-

---

[1] `https://www.quora.com/`

[2] `https://www.reddit.com/`

lish a future research benchmark through experiments with various Sequence-to-Sequence baselines to automatically generate answer for a given question about COVID-19 in Vietnamese.

3. We perform error analysis and illustrate that models trained on multiple paraphrased answers tend to have better generalization than those trained using only one answer. This reflects the advantage of having multiple paraphrased answers in single-turn conversational question answering.

The rest of this paper is structured as follows. In the following section, we review the related works. In Section 3, we describe the process of building UIT-ViCoV19QA dataset in detail, including data collection, data pre-processing, paraphrasing process, and highlight its overall statistics. Section 4 is devoted to methodologies and experiment configurations. The results and benchmarks, as well as error analysis are described in Section 5. Finally, our conclusion and future works are presented in Section 6.

2 Related works

Question answering systems consist of the single-turn and multi-turn QA. According to (Del Tredici et al., 2021), the single-turn QA takes the questions as input and returns the output without context. Single-turn QA includes Text-based QA (SQuAD (Rajpurkar et al., 2016), RACE (Lai et al., 2017)), Visual-QA (VQA (Goyal et al., 2017)), Community-based QA (ANTIQUE (Hashemi et al., 2020)), and Knowledge-based QA (MetaQA (Zhang et al., 2018)) In contrast, the multi-turn QA take the questions as input belong with contexts such as conversation history, which is called the Conversational QA (CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018)). Additionally, to verbalize the response from the question, the ParaQA dataset (Kacupaj et al., 2021) used the paraphrasing techniques. Each question in the ParaQA dataset has at least two more answers. In particular, some questions contain eight responses.

On the other hand, many efforts to create the question answering corpora in Vietnamese such as UIT-ViQuAD (Nguyen et al., 2020), UIT-ViNewsQA (Van Nguyen et al., 2020b), ViMMRC (Van Nguyen et al., 2020a) for text-based question answering, ViVQA (Tran et al., 2021) for visual-based QA, and ViCoQA (Luu et al., 2021) for conversational QA. Our works contributes to the corpora for Vietnamese as Community-based QA dataset. With the idea from the ParaQA dataset (Kacupaj et al., 2021), we manually create up to three responses from initial answer for each question in our dataset to make the dataset verbalized.

3 Dataset

In order to publish a high-quality dataset for the research community and be able to experiment with the baseline models, we have investigated building a dataset for Vietnamese, a low-resource language in the field of study. This section describes the construction of UIT-ViCoV19QA in detail and presents some statistics from the dataset. The overview of the dataset construction workflow is illustrated in Figure 1.

3.1 Data Collection

Question-answer pairs used for the constructing of UIT-ViCoV19QA dataset are extracted from FAQ documents that are publicly available on websites of respected health care organizations in Vietnam and overseas, including The Centers for Disease Control and Prevention (CDC), United Nations Children's Fund (UNICEF), The Ministry of Health of Vietnam, Vietnam Government Portal and other trusted medical institutions. Each web page provides different topics about the COVID-19 pandemic compiled in Vietnamese and often includes similar information extracted during the crawl. The following topics are covered in our dataset: origin, outbreak, and name of the disease; spread; symptoms; prevention, treatment guidelines, and nutrition; treatment models; variants of COVID-19; vaccines and vaccination; moving between areas, entry, and travel; isolation, quarantine, lockdown and social distancing; policies and sanctions; financial support; post-COVID-19; COVID-19 in children. Disease statistics in Vietnam and worldwide are not included in the dataset.

Once appropriate and trustful sources are identified, numerous handcrafted patterns are developed

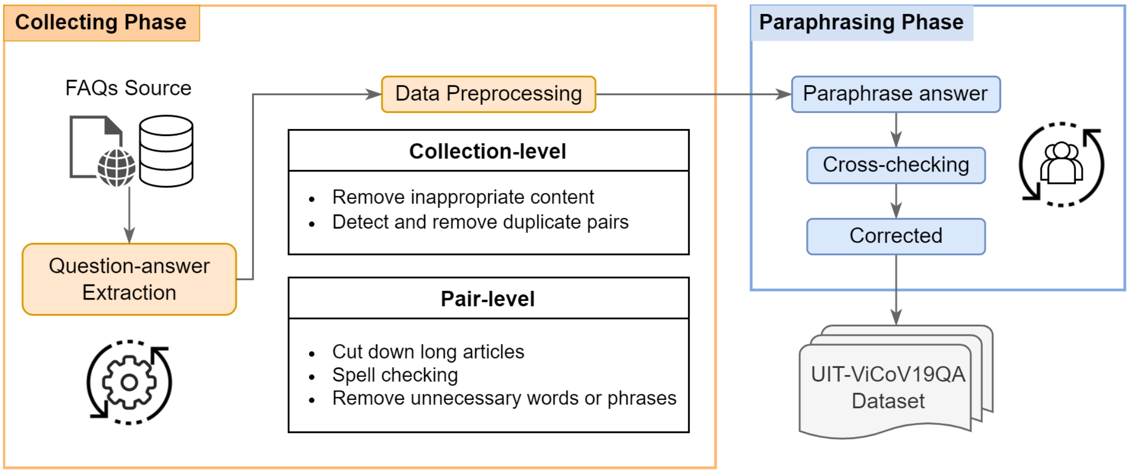

Figure 1: An overview of UIT-ViCoV19QA constructing workflow

for each website to automatically extract question-answer pairs based on the document structure, typically in HTML and PDF format. Throughout the collecting phase, the length of both question and answers are maintained to retain information. Therefore, the collected content could be a sentence, a paragraph, or a passage determined by the author or compiler. This poses a tremendous challenge for Sequence-to-Sequence models to produce good results, which makes it an ideal material for further investigations.

3.2 Data Pre-processing

A two-stage pre-processing is performed on the collected question-answer pairs. First, duplicate pairs are detected using Cosine similarity at the collection-level and are considered to be taken out of the collection. During the crawl, we also attempt to identify and remove pairs that have inappropriate content, such as articles that are neither related to COVID-19 nor the Vietnamese community or are too complicated to comprehend.

Second, at the pair-level, we check the spelling and correct mistakes. Undesirable artifacts, erroneous character and HTML tags will also be removed in this stage. In case the collected articles are too long or contains more than three paragraphs, extracted answers will be manually reduced in length by removing unnecessary information, such as source name, footnote, greetings, preliminary, over-specified explanations, and farewell. After this step, 4,500 Vietnamese question-answer pairs about COVID-19 are achieved from trusted FAQ sources. Some examples of the collected question-answer pairs are shown in Table 1.

3.3 Paraphrase Generation Process

Inspired by the concept of Kacupaj et al. (2021), we have investigated extending our dataset using the following paraphrasing methods on Vietnamese samples:

- Rearrange words, phrases, or sentences in the initial answer to create new responses without changing their meaning.
- Reduce or diversify the content of the initial answer.
- Paraphrase the initial answer using synonyms and similar structures.

These methods are manually applied on the collection to create up to three individual paraphrased responses consecutively for each question-answer pair. Newly created answers will be annotated in order to indicate the minimum number of answers per question.

After creating multiple paraphrased versions of the initial answer, we perform a cross-checking process to correct the spelling mistakes and grammatical errors as well as modify and rephrase digressive answers. By the end of this phase, the UIT-ViCoV19QA dataset is entirely constructed with 4,500 question-answer pairs containing at least one

Question: Xử lý triệu chứng ho khi chăm sóc F0 tại nhà thế nào? [**English:** How to handle cough symptoms when taking care of F0 patient at home?]

Answer: Dùng thuốc giảm ho theo đơn của bác sĩ. Có thể dùng thêm các vitamin theo đơn thuốc của bác sĩ. [**English:** Take cough suppressants as prescribed by your doctor. Additional vitamins can be taken according to the doctor's prescription.]

Question: Tôi đang điều trị viêm tiết niệu và viêm dạ dày hội chứng ruột kích thích thì có tiêm vaccine Covid-19 được hay không? [**English:** I am being treated for UTIs and gastritis with irritable bowel syndrome, can I get the Covid-19 vaccine?]

Answer: Với bệnh cấp tính mà anh/chị đang mắc phải cần được điều trị ổn định trước, sức khỏe tốt, bình thường thì có thể tiêm vaccine Covid-19. [**English:** If you have an acute illness that are stably treated, if you are in good health, you can receive the Covid-19 vaccine.]

Question: Hỗ trợ hô hấp cho trẻ em nhiễm COVID-19 ở thể nặng như thế nào? [**English:** How to provide respiratory support to children with COVID-19 in severe condition?]

Answer: Thở mask có túi Hoặc: NCPAP, HPNO, NIPPV [**English:** Apply breathing mask with bag or: NCPAP, HPNO, NIPPV]

Table 1: Examples of question-answer pairs from UIT-ViCoV19QA

answer and at most four unique paraphrased answers per question.

3.4 Statistics

The statistics of the training, development, and test sets are described in Table 2. The UIT-ViCoV19QA dataset consists of 4,500 question-answer pairs in total. In the table, the average length, as well as the vocabulary size [3] of questions and answers, are also presented.

Figure 2 illustrates the distribution of 4,500 questions of UIT-ViCoV19QA based on number of answers per question. The figure shows that the dataset contains 1800 questions that have at least two answers, 700 questions have at least three answers and half of them have a maximum of four paraphrased answers.

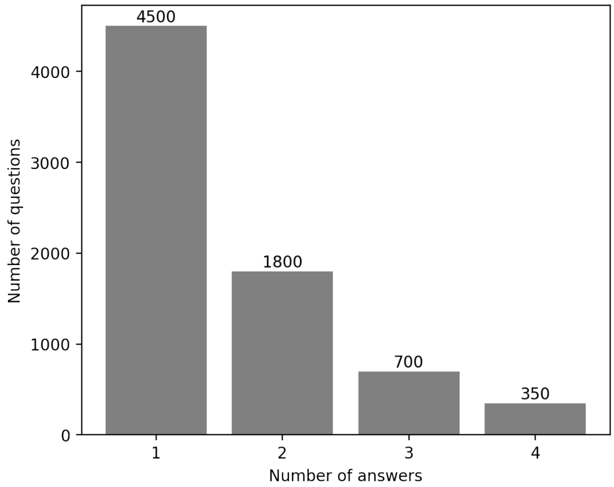

Figure 2: Distribution of total number of answers per question in UIT-ViCoV19QA

4 Methodologies

4.1 Baseline Models

This experimental section set up various deep learning models with Encoder-Decoder architecture to evaluate the UIT-ViCoV19QA dataset. These models have achieved significant results on many sequence-to-sequence learning tasks that involve long sequences, such as machine translation, text summarization, and question answering.

- **Attention-based Recurrent Neural Network**: Recurrent neural network (RNN) models used in the experiments are implemented with two attention mechanism - Bahdanau Attention (Bahdanau et al., 2016) and Luong Attention (Luong et al., 2015). To ensure that the results

---

[3] We use underthesea package: `https://github.com/undertheseanlp/underthesea` for word segmentation.

		Train	Dev.	Test	All
Answer 1	Number of question-answer pairs	3500	500	500	4500
	Average question length	31.44	33.66	32.32	31.79
	Average answer length	120.53	116.04	118.11	119.76
	Question vocabulary size	4396	1869	1770	4924
	Answer vocabulary size	8537	3689	3367	9411
Answer 2	Number of question-answer pairs	1390	209	201	1800
	Average question length	35.56	39.22	39.72	36.45
	Average answer length	40.54	39.25	42.73	40.64
	Question vocabulary size	2883	1269	1207	3305
	Answer vocabulary size	2632	1098	1129	2949
Answer 3	Number of question-answer pairs	542	79	79	700
	Average question length	34.77	36.70	39.28	35.49
	Average answer length	28.68	26.43	30.89	28.67
	Question vocabulary size	1836	717	693	2111
	Answer vocabulary size	1554	503	585	1753
Answer 4	Number of question-answer pairs	272	39	39	350
	Average question length	36.57	37.59	42.15	37.10
	Average answer length	29.75	29.03	35.72	30.25
	Question vocabulary size	1315	470	460	1519
	Answer vocabulary size	924	353	374	1075

Table 2: Overall statistics of the UIT-ViCoV19QA dataset.

are comparable, these models are set up with similar hyperparameters in the encoder and decoder as follows: an embedding layer with dimension 512, two hidden layers of 512 gated recurrent unit (GRU) cells, and a drop-out rate of 0.5. Bidirectional gated recurrent unit (Bi-GRU) is applied in the encoder of both models to help them understand the context better. For simplicity, RNN models using Bahdanau attention and Luong attention are annotated as RNN-1 and RNN-2, respectively.

- **Convolutional Network** (Gehring et al., 2017): Different from RNN, convolutional neural network uses many convolutional layers typically applied in image processing. Each layer uses filters to learn to extract different features from the text. In our experiments, the hyperparameters of models are set as follows: embedding layer with dimension 512, three convolutional layers with hidden size 512 use 1024 filters with kernel size 3 x 3, and drop-out probability 0.5.

- **Transformer** (Vaswani et al., 2017): As a dominant architecture in natural language processing (NLP), the model and its variants, such as BERT and pre-trained versions of BERT, have been commonly used to achieve state-of-the-art results for many tasks in the field. The model is set up with these settings: embedding layer with dimension 512, two layers with 8 self-attention heads, positional embedding layer with max length 500, position-wise feed-forward layer with dimension 2048 and drop-out rate 0.5.

4.2 Evaluation Metrics

Three standard metrics utilized for evaluating baseline models. These metrics are commonly used in machine translation and text summarization tasks to compare the generated text with human performance.

- **BLEU** (Papineni et al., 2002): BLEU is an n-gram based evaluation metric, widely used for Machine Translation (MT) evaluation to claim a high correlation with human judgments of quality. It aims to count the n-gram overlaps in the reference by taking the maximum count of each n-gram and clipping the count of the n-grams in the candidate text to the maximum count in the reference. In our experiments, BLEU score is calculated using unigram (BLEU-1) and 4-gram (BLEU-4) with uniform weight $w_n = 0.25$.

- **METEOR** (Lavie and Agarwal, 2007): The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. Moreover, it has several features not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. Unfortunately, those features are not yet supported for the Vietnamese language. We set up standard METEOR only using exact matching for evaluation with $\alpha = 0.9$, $\beta = 3.0$ and $\gamma = 0.5$.

- **ROUGE-L** (Lin, 2004): This metric measures the longest common subsequence (LCS) between our model output and reference. The idea here is that a longer shared sequence would indicate more similarity between the two sequences. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence-level word order. For each generated answer, we choose the best score achieved from comparing it with all existing reference answers.

4.3 Experimental Configuration

To assess the dataset's quality and illustrate the effect of having multiple paraphrased answer, we conduct and run individual experiment with four different dataset settings: one answer, two answers, three answers and finally, four paraphrased answers per question.

For Transformer, Adam optimizer is implemented with parameters as follows: $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and $warmup\_step = 2000$. This setup varies the learning rate of the model during training progress by increasing it linearly for the first 2000 training steps and decreasing it after that proportionally to the inverse square root of the step number. For other models, standard Adam optimizer with a fixed learning rate of 0.001 is applied in the training.

The training progress is configured with a batch size of 8 in 30 epochs on an NVIDIA Tesla P100 GPU via the Kaggle platform[4]. After each epoch, the performance loss on the train and development sets is calculated using the Cross-Entropy Loss function. The maximum length of the model-generated output is limited to 500 tokens to reduce the generating time of repetitive loops caused by text degeneration.

5 Experimental Results

Table 3 presents the performance of our baseline models on different settings of the UIT-ViCoV19QA dataset. The final score of each metric in the table is achieved by calculating the average scores of all model-generated answers.

Evaluated using BLEU-1 and BLEU-4, the performance of models tends to improve when applying more paraphrased answers in the experiments, though this trend behaves differently among models. RNN-1 using two answers per question achieved the best BLEU-1 of 26.62% while on BLEU-4, Transformer using four answers outperforms other models with a score of 14.38%. Apart from RNN-1 achieves best score when applying two answers, other models have highest performance when training and evaluating using four answers.

In contrast with the BLEU score, the performance of models evaluated by METEOR and ROUGE-L varies significantly under different dataset settings. As shown in the table, RNN-2 and Convolutional perform well when using only one response, with scores of 33.95% and 32.29% respectively, while RNN-1 and Transformer need to apply more answers to achieve better scores.

Three examples of models generated responses are shown in Table 4 to illustrate the generation performance comparing with the original answer.

5.1 Error Analysis

From the experiment results, we determine that the Transformer trained on four answers gives the best

---

[4] `https://www.kaggle.com/`

Model	# answers	BLEU-1 (%)	BLEU-4 (%)	METEOR (%)	ROUGE-L (%)
RNN-1 (Bahdanau et al., 2016)	1	21.79	10.29	25.34	32.36
	2	**26.62**	**12.86**	25.15	33.66
	3	26.09	12.63	**25.98**	**33.68**
	4	24.56	12.27	23.91	32.57
RNN-2 (Luong et al., 2015)	1	21.04	10.94	**24.72**	**33.95**
	2	24.07	11.30	23.81	31.92
	3	23.8	10.95	23.65	31.65
	4	**24.38**	**12.30**	24.29	32.09
Convolutional (Gehring et al., 2017)	1	19.17	9.51	**22.76**	**32.29**
	2	21.63	10.81	20.66	30.71
	3	22.08	10.68	20.96	30.84
	4	**23.26**	**10.93**	21.84	31.15
Transformer (Vaswani et al., 2017)	1	21.84	10.82	23.37	31.67
	2	24.81	13.22	**24.60**	**32.25**
	3	24.90	13.64	21.72	32.20
	4	**25.19**	**14.38**	23.11	32.19

Table 3: Performance of baseline models on different dataset settings of UIT-ViCoV19QA

performance among others based on BLEU-4 according to the approach of Kacupaj et al. (2021). Twenty samples are randomly chosen from the generated answers of the model to perform error analysis. We calculate the average length and vocabulary size and count the number of POS tags in the references and generated answers. The statistics of the analysis process are shown in Table 5.

	Original answers	Generated answers
Avg. length	115.53	75.25
Vocabulary size	776	114
# POS tag	16	12

Table 5: The average length, vocabulary size and number of POS tag in 20 generated samples and their references

The vocabulary size of the output answers is almost seven times less than that of the original answers, while the average length and number of POS tags do not vary significantly. This implies that Transformer does not generalize well, and the tokens overlapping or text degeneration may have occurred in the outputs of the model.

By checking the generated results of Transformer, we notice some output answers contain overlapping phrases and do not match the question and reference answers. Further review the output of other models, various mistakes and errors have also been pointed out in the generated responses. In summary, these errors can be divided into three main groups.

- The first group includes meaningless and illogical answers. In some cases, the generated answers are just a set of tokens arranged in a chaotic order.
- The second group contains meaningful and coherent answers but do not satisfy the question's requirements, as illustrated by the second example in Table 4. This type of error can be caused by the diverse Vietnamese vocabulary and grammar or long passage, which makes the model difficult to understand the context of question.
- The third group consists of text degeneration. Typically, this is a common phenomenon in sequence-to-sequence learning tasks when a word or a phrase is infinitely repeated in the generated sequence indicating that the models may not generalize well. As illustrated by the third example in Table 4, the generated response does not satisfy the question

Question: tôi bị dị ứng với thuốc giảm đau giãn cơ . tôi uống thuốc này lúc bị sút lưng (phồng đĩa đệm) . vậy xin hỏi tôi có tiêm vaccine covid-19 được không ạ ? [**English:** I am allergic to muscle relaxants. I took this medicine when I had a low back pain (bulging disc). So can I get the covid-19 vaccine?]

Original Answer: chào anh / chị , nếu tình trạng dị ứng của anh / chị ở mức độ thông thường , không có phản vệ từ độ 2 trở lên , anh / chị vẫn có thể tiến hành tiêm chủng tại bệnh viện được . chúc anh / chị nhiều sức khỏe . trân trọng ! [**English:** hello sir / madam , if your allergy is at normal level , not having 2nd degree anaphylaxis or higher , you can still carry out vaccination at the hospital . Wish you a lot of health. Best regards !]

Generated Answer: chào anh , theo hướng dẫn của bộ y tế , anh thuộc nhóm đối tượng cần thận trọng khi tiêm chủng vaccine covid-19 . tuy nhiên , anh cần cung cấp đầy đủ thông tin về tình trạng dị ứng của mình cho bác sĩ khám sàng lọc để được tư vấn cụ thể hơn . cảm ơn câu hỏi của anh . trân trọng ! [**English:** hello sir, according to the guidance of the Ministry of Health, you are in the group of subjects that need to be cautious when vaccinating against the covid-19 vaccine. however, you need to provide enough information about your allergy to the screening doctor for more specific advice. Thanks for your question. best regards !]

Question: ai không nên tiêm vaccine covid-19 ? [**English:** who should not get the covid-19 vaccine?]

Original Answer: theo tổ chức y tế thế giới (who) , mỗi quốc gia cần chủng ngừa cho ít nhất 70 % dân số mới đạt được miễn dịch cộng đồng . giống với các loại dược phẩm khác , vaccine covid-19 có thể gây một số tác dụng phụ và không phù hợp ở từng đối tượng nhất định . tất cả những người có phản ứng dị ứng nghiêm trọng với bất kỳ thành phần hoạt chất hoặc tá dược nào liệt kê trong mục thành phần vắc xin đều không được tiêm vắc xin . bên cạnh đó , cần thận trọng tiêm chủng với các đối tượng , như : nếu bạn đã từng có phản ứng dị ứng nghiêm trọng (phản ứng phản vệ) sau khi tiêm vắc xin nào đó trước đây . nếu hệ miễn dịch của bạn bị suy yếu (suy giảm miễn dịch) hoặc bạn đang dùng thuốc làm suy yếu hệ miễn dịch (như corticosteroid liều cao , thuốc ức chế miễn dịch hoặc thuốc ung thư) . nếu bạn hiện đang bị nhiễm trùng nặng với thân nhiệt cao (trên 38 °c / 100.4 °f) . nếu bạn có vấn đề về xuất huyết / chảy máu hoặc bầm tím , hoặc nếu bạn đang dùng thuốc làm loãng máu (thuốc chống đông máu) . trong trường hợp nếu bạn không chắc chắn bất kỳ điều gì bên trên , hãy trao đổi với bác sĩ hoặc chuyên gia chăm sóc sức khỏe của bạn trước khi bạn được tiêm vắc xin . [**English:** according to the world health organization (who) , every country needs to immunize at least 70 % of the population to achieve herd immunity . Like other medicines, the covid-19 vaccine can cause some side effects and is not suitable for certain people. All people who have had a severe allergic reaction to any of the active ingredients or excipients listed in the vaccine ingredients section should not get the vaccine. In addition , care should be taken in vaccinating subjects , such as : if you have ever had a severe allergic reaction (anaphylaxis) after receiving a certain vaccine in the past . if your immune system is weakened (immunocompromised) or you are taking medicine that weakens your immune system (such as high-dose corticosteroids , immunosuppressants or cancer drugs) . if you currently have a severe infection with a high body temperature (above 38 °c / 100.4 °f) . if you have bleeding / bleeding or bruising problems , or if you are taking blood thinners (anticoagulants) . In case if you are not sure of any of the above, talk to your doctor or healthcare professional before you receive the vaccine.]

Generated Answer: người dân cần thực hiện nghiêm túc , chấp hành quy định của địa phương , thực hiện thông điệp 5k để bảo vệ bản thân và những người xung quanh .[**English:** people need to strictly comply with local regulations and implement the 5k message to protect themselves and those around them.]

Question: biến chủng Epsilon (B. 1.427 / B. 1.429) là gì ? [**English:** what is Epsilon (B. 1,427/B. 1,429) variant?]

Original Answer: biến chủng này có thể tự sắp xếp lại một phần của protein đột biến mà virus sử dụng để liên kết với các tế bào của cơ thể . vì thế , biến chủng epsilon có khả năng lây nhiễm cao hơn khoảng 20 % và dễ dàng làm suy yếu hệ miễn dịch của con người . [**English:** this variant can rearrange a portion of the mutated protein that the virus uses to bind to the body's cells. Therefore, epsilon variants are about 20% more contagious and easily weaken the human immune system.]

Generated Answer: theo quy định tại điều 11 nghị định số 176 / 2013 / nđ-cp ngày 14/11/2013 của chính phủ về xử phạt hành chính trong lĩnh vực y tế , hành vi không thực hiện biện pháp cách ly y tế tại điều 11 nghị định số 176 / 2013 / nđ-cp ngày 14/11/2013 của chính phủ về xử phạt hành chính trong lĩnh vực y tế , hành vi có hành vi có hành vi này bị phạt tiền từ 5.000.000 đồng đến 10.000.000 đồng đến 10.000.000 đồng đến 10.000.000 đồng ... [**English:** in accordance with Article 11 of the Government's Decree No. 176/2013/nd-cp dated November 14, 2013 on administrative sanctions in the health sector, failure to implement medical isolation measures in Article 11 of the Government's Decree No. 176/2013/nd-cp dated November 14, 2013 on administrative sanctions in the health sector, acts committed in this act are subject to a fine of from VND 5,000,000,000 to VND 10,000,000 to VND 10,000,000 to VND 10,000,000 ...]

Table 4: Examples of answers generated by the proposed models compared with the original answers

and the phrase "tới 10.000.000 đồng"(to VND 10,000,000) keeps repeated until the end of the answer.

6 Conclusion and Future Works

In this paper, we presented UIT-ViCoV19QA, the first community-based question answering dataset about COVID-19 for Vietnamese. Our dataset comprises 4,500 question-answer pairs with multiple paraphrased answers. The dataset's quality was evaluated through various baseline deep learning models and commonly used metrics such as BLEU, METEOR, and ROUGE-L.

We illustrated the effect of having multiple paraphrased answers for experiments with baseline models and provided benchmark results for further research. RNN with Bahdanau attention achieves the best BLEU-1 and METEOR scores of 26.62% and 25.98% when applying two and three answers respectively. Transformers using four answers outperforms others on BLEU-4 with score of 14.38%. On ROUGE-L, RNN with Luong Attention using one answer has the best performance of 33.95%. The advantage of having multiple paraphrased answers is greatly illustrated by BLEU scores, on which three out of four models achieve the best performance when applying all four paraphrased answers. On the contrary, our experiments showed that METEOR and ROUGE-L scores do not give a clear reflection of the improvement in models performance when increase number of answer used.

Through error analysis, we showed that the performance of these models is not quite good since the generated answers contain various errors. There are several reasons for this: the diversity of the Vietnamese language, lacking a specific evaluation metric for the Vietnamese language, the long-sequence content, and the size limitation of our dataset. The dataset offers a valuable contribution to the community, providing the foundation for many research lines in the single-turn QA domain and other areas.

In the future, UIT-ViCoV19QA can be expanded in size by collecting more relevant question-answer pairs and creating more paraphrased answers. The embedding layer of the proposed baselines can also be investigated to be replaced with pre-trained word embeddings for Vietnamese, such as PhoBERT (Nguyen and Nguyen, 2020), to improve model performance.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[Bahdanau et al.2016] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

[Choi et al.2018] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

[Del Tredici et al.2021] Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and Adrià de Gispert. 2021. Question rewriting for open-domain conversational qa: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2974–2978.

[Gehring et al.2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

[Goyal et al.2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

[Hashemi et al.2020] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer.

[Kacupaj et al.2021] Endri Kacupaj, Barshana Banerjee, Kuldeep Singh, and Jens Lehmann. 2021. Paraqa: A question answering dataset with paraphrase responses for single-turn conversation.

[Lai et al.2017] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference* *on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Lavie and Agarwal2007] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June. Association for Computational Linguistics.

[Lin2004] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

[Luong et al.2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

[Luu et al.2021] Son T Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. In *International Conference on Computational Collective Intelligence*, pages 546–558. Springer.

[Nguyen and Nguyen2020] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

[Nguyen et al.2020] Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[Qiu and Huang2015] Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Twenty-Fourth international joint conference on artificial intelligence*.

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

[Reddy et al.2019] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

[Tran et al.2021] Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, and Kiet Van Nguyen. 2021. Vivqa: Vietnamese visual question answering. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 546–554.

[Van Nguyen et al.2020a] Kiet Van Nguyen, Khiem Vinh Tran, Son T Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020a. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

[Van Nguyen et al.2020b] Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. New vietnamese corpus for machine reading comprehension of health news articles. *arXiv preprint arXiv:2006.11138*.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Zhang et al.2018] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Thirty-second AAAI conference on artificial intelligence*.

[Zhang et al.2021] Wei Zhang, Zeyuan Chen, Chao Dong, Wen Wang, Hongyuan Zha, and Jianyong Wang. 2021. Graph-based tri-attention network for answer ranking in cqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14463–14471.